Encoding – Eine Geschichte voller Missverst�ndnisse

Felix Dreißig

5. November 2019

FSI Informatik Lightning Talks #1

Stand 2019

```
Folien.tex:36: Package inputenc Error: Unicode
character ﬀ (U+FFFD)
Folien.tex:36: leading text: \maketitle
```

Stand 2019

```
Folien.tex:36: Package inputenc Error: Unicode
character ﬀ (U+FFFD)
Folien.tex:36: leading text: \maketitle
```

Die Anfänge

- Morse, Baudot etc. (19. Jahrhundert)
- BCDIC 1928, ASCII ~ 1960, EBCDIC 1963

```
> man ascii
```

```
     2 3 4 5 6 7
    ------------
  0:   0 @ P ` p
  1: ! 1 A Q a q
  2: " 2 B R b r
  3: # 3 C S c s
  4: $ 4 D T d t
  5: % 5 E U e u
  6: & 6 F V f v
  7: ' 7 G W g w
  8: ( 8 H X h x
  9: ) 9 I Y i y
  A: * : J Z j z
  B: + ; K [ k {
  C: , < L \ l |
  D: - = M ] m }
  E: . > N ^ n ~
  F: / ? O _ o DEL
```

```
     30 40 50 60 70 80 90 100 110 120
    ---------------------------------
  0:    (  2  <  F  P  Z  d   n   x
  1:    )  3  =  G  Q  [  e   o   y
  2:    *  4  >  H  R  \  f   p   z
  3: !  +  5  ?  I  S  ]  g   q   {
  4: "  ,  6  @  J  T  ^  h   r   |
  5: #  -  7  A  K  U  _  i   s   }
  6: $  .  8  B  L  V  `  j   t   ~
  7: %  /  9  C  M  W  a  k   u  DEL
  8: &  0  :  D  N  X  b  l   v
  9: '  1  ;  E  O  Y  c  m   w
```

ISO-8859 (ab 1987)

Für Westeuropa: ISO-8859-1 (Latin-1)

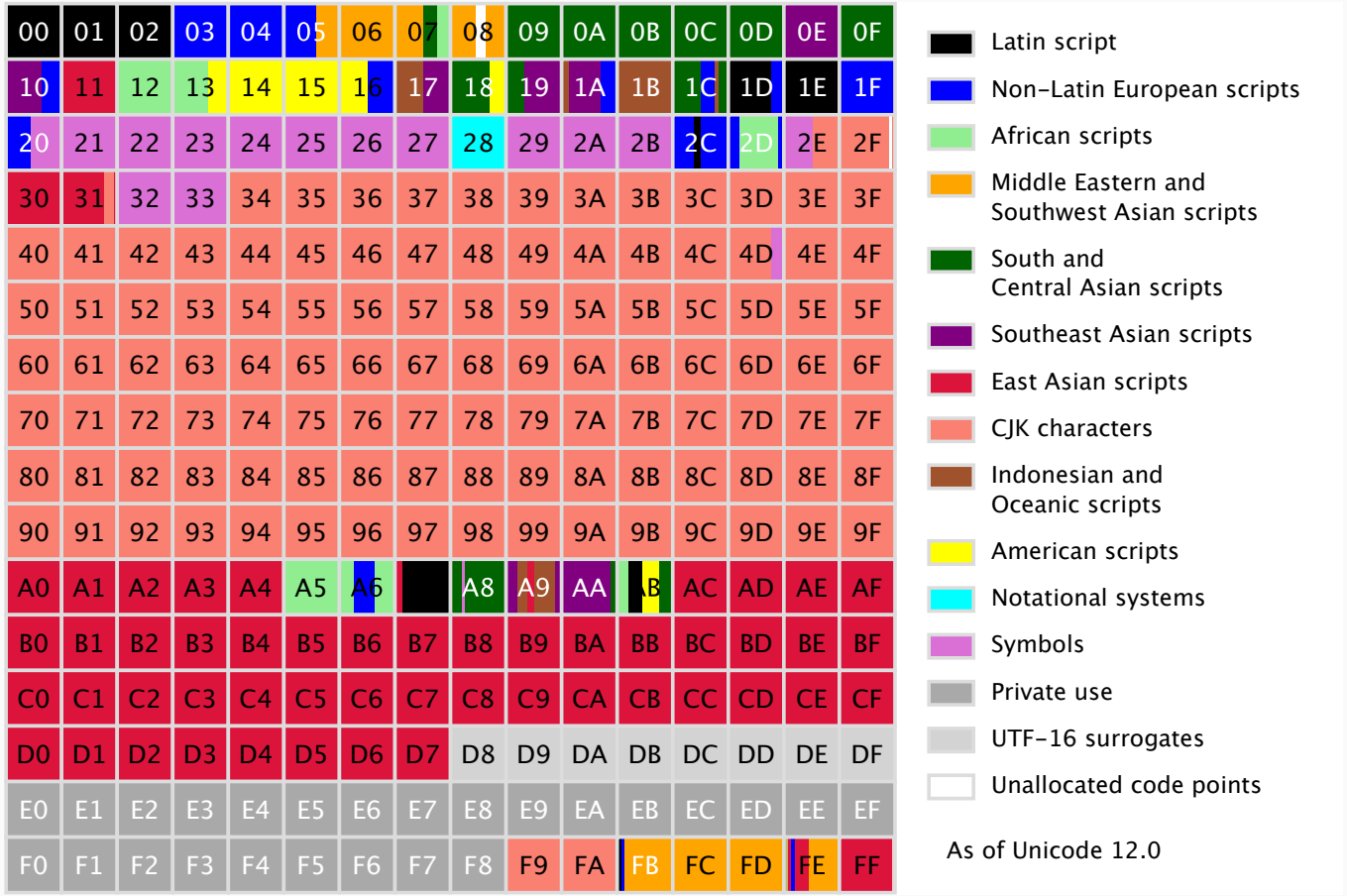
7_ 112	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	
8_ 128																
9_ 144																
A_ 160	NBSP 00A0	¡ 00A1	¢ 00A2	£ 00A3	¤ 00A4	¥ 00A5	¦ 00A6	§ 00A7	¨ 00A8	© 00A9	ª 00AA	« 00AB	¬ 00AC	SHY 00AD	® 00AE	¯ 00AF
B_ 176	° 00B0	± 00B1	² 00B2	³ 00B3	´ 00B4	µ 00B5	¶ 00B6	· 00B7	¸ 00B8	¹ 00B9	º 00BA	» 00BB	¼ 00BC	½ 00BD	¾ 00BE	¿ 00BF
C_ 192	À 00C0	Á 00C1	Â 00C2	Ã 00C3	Ä 00C4	Å 00C5	Æ 00C6	Ç 00C7	È 00C8	É 00C9	Ê 00CA	Ë 00CB	Ì 00CC	Í 00CD	Î 00CE	Ï 00CF
D_ 208	Ð 00D0	Ñ 00D1	Ò 00D2	Ó 00D3	Ô 00D4	Õ 00D5	Ö 00D6	× 00D7	Ø 00D8	Ù 00D9	Ú 00DA	Û 00DB	Ü 00DC	Ý 00DD	Þ 00DE	ß 00DF
E_ 224	à 00E0	á 00E1	â 00E2	ã 00E3	ä 00E4	å 00E5	æ 00E6	ç 00E7	è 00E8	é 00E9	ê 00EA	ë 00EB	ì 00EC	í 00ED	î 00EE	ï 00EF
F_ 240	ð 00F0	ñ 00F1	ò 00F2	ó 00F3	ô 00F4	õ 00F5	ö 00F6	÷ 00F7	ø 00F8	ù 00F9	ú 00FA	û 00FB	ü 00FC	ý 00FD	þ 00FE	ÿ 00FF

Unicode (ab 1991)

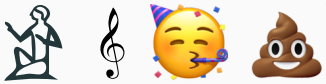
- Große Tabelle
- Ursprünglich 2^{16} Zeichen
- Heute 1.114.112 *Code points* in 17 *Planes* à 65.536 (2^{16})
- Belegt: Ca. 138.000 *Code points*
- Zusätzlich: *Composite characters*

Was ist drin im Unicode?

Plane 0: *Basic Multilingual Plane*



Plane 1: *Supplementary Multilingual Plane*



UCS und UTF-16

- UCS-2: 16 Bit fix $\rightarrow$ Nur *Basic Multilingual Plane*
- UTF-16: Erweiterung von UCS-2, Zeichen außerhalb der BMP werden durch *Surrogate* (2 x 16 Bit) codiert
- UCS-4: 32 Bit fix

UTF-8: Geschichte

```
Subject: UTF-8 history
From: "Rob 'Commander' Pike" <r (at) google.com>
Date: Wed, 30 Apr 2003 22:32:32 -0700 (Thu 06:32 BST)
To: mkuhn (at) acm.org, henry (at) spsystems.net
Cc: ken (at) entrisphere.com

Looking around at some UTF-8 background, I see the same incorrect
story being repeated over and over. The incorrect version is:
    1. IBM designed UTF-8.
    2. Plan 9 implemented it.
That's not true. UTF-8 was designed, in front of my eyes, on a
placemat in a New Jersey diner one night in September or so 1992.
```

https://www.cl.cam.ac.uk/~mgk25/ucs/utf-8-history.txt

UTF-8: Funktionsweise

Number of bytes	Bits for code point	First code point	Last code point	Byte 1	Byte 2	Byte 3	Byte 4
1	7	U+0000	U+007F	`0xxxxxxx`			
2	11	U+0080	U+07FF	`110xxxxx`	`10xxxxxx`		
3	16	U+0800	U+FFFF	`1110xxxx`	`10xxxxxx`	`10xxxxxx`	
4	21	U+10000	U+10FFFF	`11110xxx`	`10xxxxxx`	`10xxxxxx`	`10xxxxxx`

- Abwärtskompatibel mit ASCII
- Für viele Alphabete relativ effizient
- Selbstsynchronisierend

Encoding erkennen

- Heuristiken

```
> file -i foo.txt
foo.txt: text/plain; charset=utf-8
```

- BOM
- Spezifikation
- Auszeichnung

```
<meta charset="UTF-8" />
```

Praktische Tipps

- Verwendet UTF-8 zum Speichern und Übertragen
- Nutzt intern die Unicode-Strings eurer Programmiersprache
- Baut *Unicode-Sandwiches*

Encoding – Eine Geschichte voller Missverst�ndnisse

Felix Dreißig

5. November 2019

FSI Informatik Lightning Talks #1